HIGH CARDINALITY ATTRIBUTES IN THE DATASET—

→ Many times we have nominal variables with many distinct values. For example ZIP code or street name of a person, the university or high school. In United State for example, there are approximately 43,000 zip codes. Such features could be very predictive as it could be telling that someone lives in a certain village or works in certain business sector.

→ We call such nominal variable with more than 1000 distinct values "high cardinality attributes".

→ Despite their potential, unfortunately such variables are typically discarded. Firstly, including these attributes by standard dummy encoding increasing dimensionality of data to such extent, model unable to process them which also leads to thousand or even millions of features. Secondly, for some time some variable can try to group value in semantic manner, like grouping on state level or first two digit etc but not possible every time.

Methods to address this issue →

1) **Supervised ratio** — The easiest transformation is transforming each ZIP code to percentage of positive instances in ZIP code. When predicting about FIXED DEPOSIT Customers, In ZIP code 411015 will be given a transformed value of the percentage of FD takers in that ZIP code. So if training set contains 100 customers, 5 of which are from 411015 then transformed value is 0.05.

$$\text{Supervised Ratio} = \frac{\text{FD takeup at Pincode}}{\text{Total FD takeup}}$$

2) **WOE (Weight of Evidence)** — This transformation can be used in credit scoring for low value nominal in the past.

$$WOE = \ln\left(\frac{P_i^\circ / TP}{N_i^\circ / TN}\right)$$

$P_i^\circ \rightarrow 0.05$
$W_i^\circ \rightarrow 0.95$
$TP \rightarrow$ True Positive
$TN \rightarrow$ True Negative.